Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference

Supplementary Material

7. Related Work

Interpretability of LLMs. Research on attention mechanisms has significantly enhanced our understanding of large language models. For instance, Xiao et al. [42] highlight a phenomenon known as *attention sink*, indicating that maintaining the key-value states of initial tokens can largely restore the performance of window attention, primarily due to the strong attention scores associated with these tokens. Furthermore, Wang et al. [40] discovered that label words serve as anchors in in-context learning, facilitating the aggregation and distribution of task-relevant information. In addition, Wu et al. [41] identified a specific category of attention heads, referred to as retrieval heads, which are primarily responsible for extracting relevant information from lengthy contexts. However, most studies on attention mechanisms focus exclusively on text-based models, creating a gap in our understanding of information interaction within MLLMs. Our research aims to bridge this gap, offering new insights into how MLLMs process and utilize visual information.

Inference Optimization for LLMs. Research on efficient inference in large language models has primarily focused on two categories of optimization: (1) Memory Consumption Optimization, which includes methods such as FlashAttention [9], vLLM [13], and RingAttention [20] that enhance the memory efficiency of the attention module without significantly altering outcomes; and (2) Computation Simplification, which involves techniques like StreamingLLM and FastGen [11] that improve inference efficiency by eliminating redundant attention calculations. This paper emphasizes the latter category. Most existing methods target text-only models, creating a notable gap in their applicability to MLLMs. Recent strategies, including FastV and VTW [19], have accelerated inference speeds through image token pruning, yet they overlook the shift in the dominant flow of visual information, failing to fully harness the potential for accelerating the inference of MLLMs.

8. Results of modality impact assessment

Fig. 9 illustrates the influence of various modalities on the prediction outcomes of the LLaVA-1.5-7B and LLaVA-1.5-13B models within the Sci-VQA and AOKVQA datasets.

9. Results of visual flow analysis

Fig. 10 illustrates the significance of intra-visual flow compared to visual-textual flow in the LLaVA-1.5-7B and LLaVA-1.5-13B models within the Sci-VQA and AOKVQA datasets. In shallow layers, the importance of *visual-textual information flow* is notably high, while *intra-visual information flow* is comparatively low. In deeper layers, *intra-visual information flow* becomes dominant.

10. Details for computation of prediction bias

The prediction biases, $E_{vv,l}$ and $E_{vt,l}$, resulting from disruptions in *visual-textual* and *intra-visual information flows*, as introduced in Sec. 3.3, may have caused confusion. Here, we provide a more detailed explanation of their calculation methods.

In the absence of disruption to information flow, **Score Consistency** of the model is denoted as C. When *intra-visual information flow* in the l-th layer is sidrupted, **Score Consistency** is represented as $C_{vv,l}$. The prediction bias $E_{vv,l}$ resulting from this disruption is calculated as follows:

$$E_{vv,l} = C - C_{vv,l}. \tag{11}$$

Similarly, **Score Consistency** of the model after disrupting *visual-textual information flow* in the l-th layer is denoted as $C_{vt,l}$. Consequently, the prediction bias $E_{vt,l}$ resulting from this disruption is calculated as follows:

$$E_{vt,l} = C - C_{vt,l}. \tag{12}$$

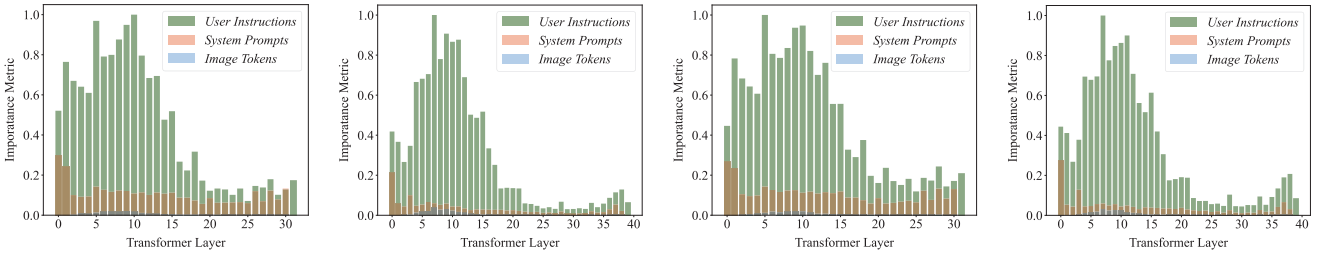
11. Reasons for Using Bias Ratio

We use the D_l metric to validate the importance of the intra-visual information flow, based on two main considerations:

- As demonstrated by the experimental results in Sec. 2.2, the prediction outcomes are primarily influenced by intra-textual information flow, which weakens as the network depth increases. Consequently, although intra-visual information flow becomes more prominent in deeper layers, its disruption has minimal impact on prediction outcomes. Therefore, we use the significance of visual-textual information flow as a baseline and apply a logarithmic ratio to measure the variation in the importance of intra-visual information flow.
- We focus on the relative strength between intra-visual and visual-textual information flows to clearly illustrate the shift in the mechanism of visual information processing in Multimodal large language models.

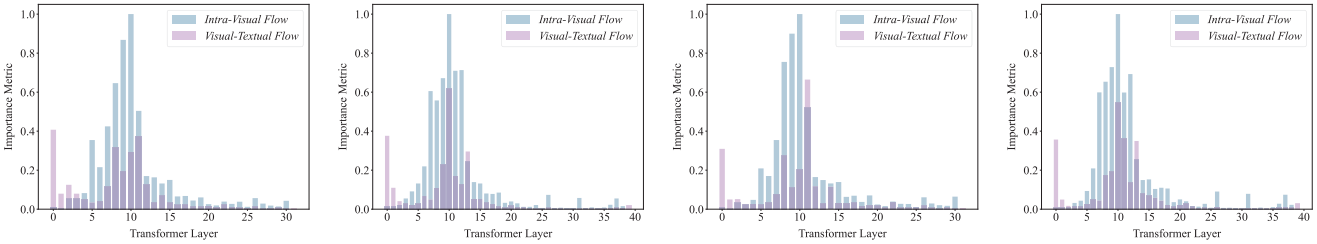
12. Experimental Results for HiMAP

Section 12.1 discusses the performance on ChartQA and DocQA datasets, Section 12.2 presents results on the MME



(a) Sci-VQA (LLaVA-7B) (b) Sci-VQA (LLaVA-13B) (c) A-OKVQA (LLaVA-7B) (d) A-OKVQA (LLaVA-13B)

Figure 9. Experimental Results of modality impact assessment. The contribution of visual modality is lower than textual modality.



(a) Sci-VQA (LLaVA-7B) (b) Sci-VQA (LLaVA-13B) (c) A-OKVQA (LLaVA-7B) (d) A-OKVQA (LLaVA-13B)

Figure 10. Additional Experimental Results of visual flow analysis. Dominant flow of visual information shifts as model depth increases.

Benchmark, Section 12.3 analyzes HiMAP's impact on content generation in LLaVA-Bench, and Section 12.4 compares the inference speed improvements between HiMAP and FastV.

Model	Method	Ratio	ChartQA	DocQA
LLaVA-7B	Baseline	100%	**9.7**	8.6
	HiMAP	**24%**	9.4	**8.8**
QwenVL-7B	Baseline	100%	65	64.9
	HiMAP	**23%**	**65.1**	**65.3**

Table 5. **Performance on ChartQA and DocQA datasets.** In each configuration, the **highest scores** are highlighted in **red**, while the **lowest computational cost** are marked in **green**. The parameters for HiMAP are set as $K_1 = 2$, $R_1 = 50\%$, $K_2 = 8$, and $R_2 = 75\%$.

12.1. Results on ChartQA & Doc-QA

We conducted a comprehensive evaluation of HiMAP's performance on the ChartQA [25] and DocQA [26] datasets, utilizing the LLaVA-v1.5 model family as our foundation. The experimental results, summarized in Tab. 5, demonstrate that effectively reduces computational overhead with minimal loss in model performance. This highlights HiMAP's efficacy in fine-grained visual question-answering tasks, showing that its pruning of visual tokens does not compromise the model's ability to perceive image details.

12.2. Results on MME

Tab. 6 illustrates the experimental outcomes of LLaVA-v1.5-7B model on the MME benchmark after incorporating HiMAP. The results indicate that, for both perception-focused and cognition-focused tasks, HiMAP method not only significantly reduces computational costs but also preserves or marginally enhances the model's performance.

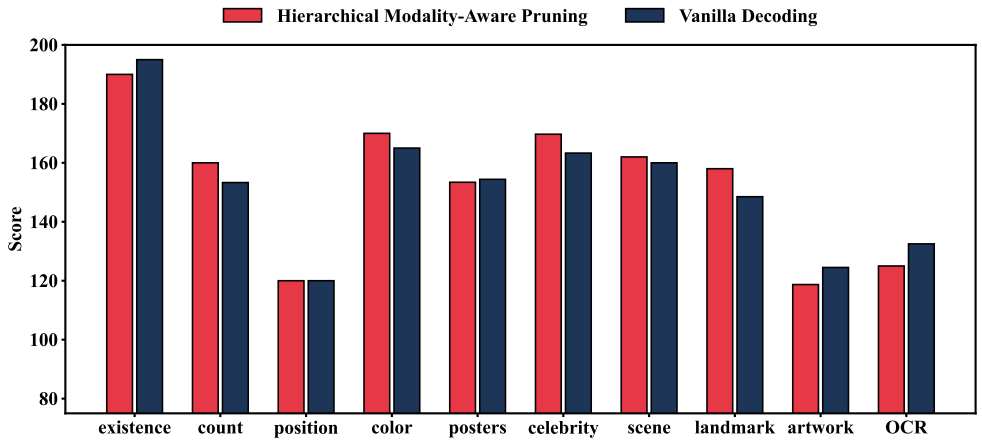
Model	Method	Ratio	MME-P	MME-C
LLaVA-7B	Baseline	100%	1459.2	290.7
	HiMAP	**24%**	**1492.6**	**292.5**
LLaVA-13B	Baseline	100%	1517.3	277.1
	HiMAP	**23%**	**1526.9**	**282.5**

Table 6. **Performance on MME Benchmark.** In each configuration, the **highest scores** are highlighted in **red**, while the **lowest computational cost** are marked in **green**. The parameters for HiMAP are set as $K_1 = 2$, $R_1 = 50\%$, $K_2 = 8$, and $R_2 = 75\%$.

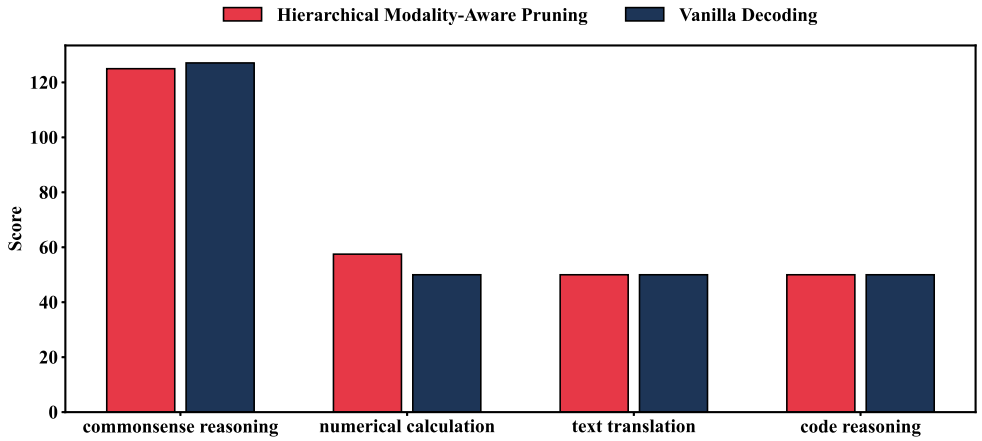
Fig. 11 highlights the performance of the LLaVA-v1.5-7B model on each subtask of the MME Benchmark after applying HiMAP, demonstrating that HiMAP effectively sustains the model's performance across all subtasks.

12.3. Case Study on LLaVA-Bench

Fig. 12 illustrates the long-text generation performance of the LLaVA-v1.5-7B model on LLaVA-Bench after the appli-



(a) Perception-Related Tasks from MME Benchmark



(b) Cognition-Related Tasks from MME Benchmark

Figure 11. **Performance of LLaVA-v1.5-13B model on the MME Benchmark.** After applying HiMAP method, the model retained nearly all of its original performance across each task.

Model	Method	Accurancy	Total Time	GPU-Memory	Latency/Example
LLaVA-v1.5-7B	Baseline	67.9	6:36	17G	0.197s
	FastV	67.8	4:51	15G	0.144s
	HiMAP	68.2	**3:54**	**14G**	**0.116s**
LLaVA-v1.5-13B	Baseline	71.6	10:45	31G	0.320s
	FastV	71.2	7:13	26G	0.214s
	HiMAP	72.5	**5:13**	**23G**	**0.158s**

Table 7. **Comparison of inference speed and GPU memory usage between HiMAP and FastV.** HiMAP outperforms FastV by delivering faster inference speeds and lower GPU memory usage while maintaining higher prediction accuracy. In each configuration, the **fastest inference speed** and the **lowest GPU memory usage** are highlighted in **green**.

cation of HiMAP. The results indicate that, when HiMAP is configured with appropriate parameters, it achieves a substantial reduction in computational overhead while almost entirely retaining the model's ability to produce high-quality long-text responses. Remarkably, even with aggressive parameter settings ($K_1 = 2$, $R_1 = 50\%$, $K_2 = 8$, $R_2 = 87.5\%$), the model consistently delivers fluent and accurate outputs.

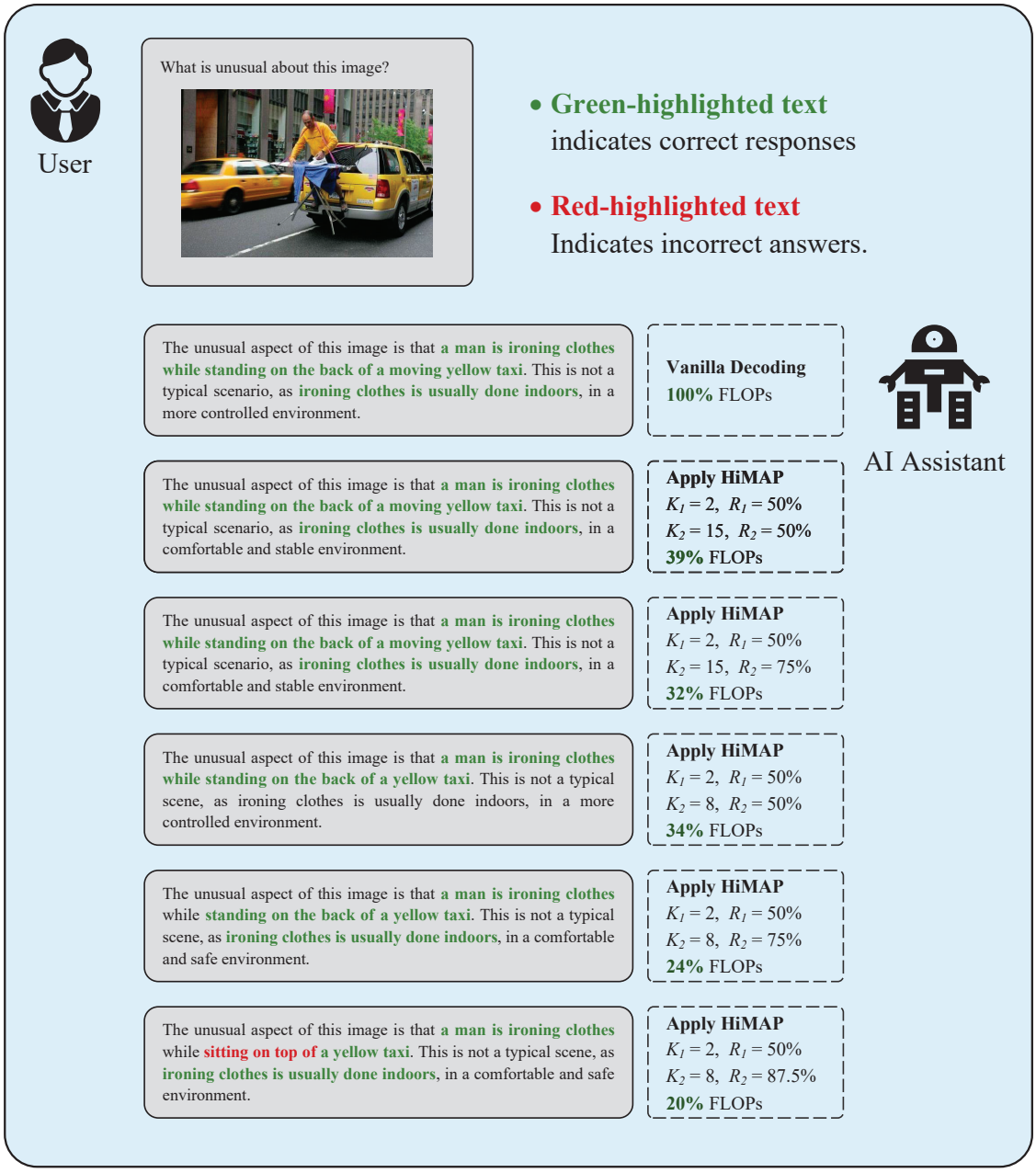


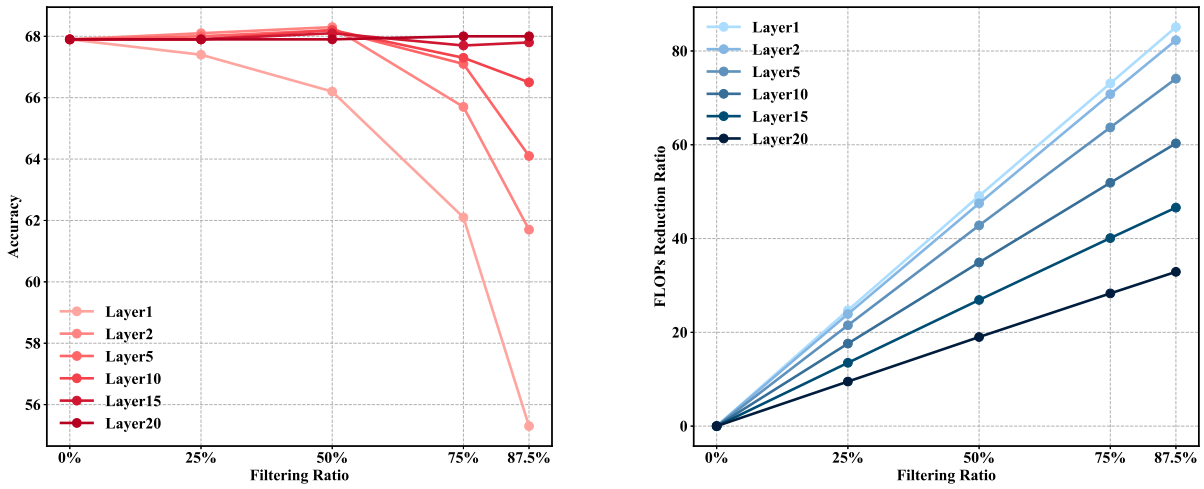
Figure 12. The output results after applying HiMAP method. **Correct segments** of outputs are highlighted in **green**, while **incorrect segments** are marked in **red**. The findings indicate that HiMAP does not compromise the quality of the responses generated by the model.

12.4. Comparison of Inference Speeds

We utilized the LLaVA-v1.5 model family to evaluate the inference speed and GPU memory usage of HiMAP and FastV on the ScienceQA dataset. The results, presented in Tab. 7, show that applying HiMAP achieves higher prediction accuracy, faster inference speed, and lower GPU memory consumption compared to FastV. These improvements are primarily driven by HiMAP's ability to perform precise and efficient pruning of visual tokens. By leveraging different vision-dominant information streams at the model's shallow and deep layers, HiMAP maximizes the potential for inference acceleration.

13. Ablation Studies on HiMAP

Sec. 13.1 delves into the impact of HiMAP's parameters, K and R, on pruning performance. Sec. 13.2 evaluates the individual contributions of the shallow-layer and deeper-layer pruning modules to predictions.



(a) Accuracy

(b) FLOPs Reduction Ratio

Figure 13. **Ablation Study on the Parameters K_1 and R_1.**

Model	K_2	R_2	TFLOPs	FLOPs Ratio	ScienceQA	A-OKVQA	NoCaps	Flickr30k
LLaVA-v1.5-7B	Baseline		2.98	100%	67.9	76.7	78.8	50.9
	8	87.5%	0.59	20%	68	71.7	74.6	46.2
	8	75%	0.73	24%	**68.3**	**77.2**	76.1	47.2
	8	50%	1.01	33%	67.8	**76.7**	76.9	49.1
	15	87.5%	0.88	29%	68.1	77.2	77.5	50.1
	15	75%	0.97	32%	68.2	77.2	**78.7**	**51.3**
	15	50%	1.17	39%	**68.2**	77.2	**79.2**	**51.7**
LLaVA-v1.5-13B	Baseline		5.81	100%	71.6	82	82.8	53.6
	8	87.5%	1.08	18%	72	79.9	77.5	47.5
	8	75%	1.36	23%	**72.1**	**81.4**	82.5	52.6
	8	50%	1.94	33%	71.9	**81.2**	82.7	52.8
	15	87.5%	1.52	26%	71.7	81	82.9	52.6
	15	75%	1.74	30%	**72.5**	81.2	**83.7**	**53.8**
	15	50%	2.19	37%	72.1	81.1	**83.9**	**54.1**

Table 8. **Ablation Study on K_2 and R_2.** In each configuration, the **highest score** is marked in **red**, while the **second-highest score** is marked in **blue**.

13.1. Effect of Filtering Layer & Filtering Ratio

Ablation studies were performed on parameters K_1 and R_1. After excluding the deeper-layer pruning module, we tuned K_1 and R_1 to assess their influence on HiMAP's pruning effectiveness. As illustrated in Fig. 13, it is clear that pruning less than 50% of visual tokens beyond the second model layer does not substantially impact prediction accuracy.

We conducted further ablation experiments on the parameters K_2 and R_2. By fixing $K_1 = 2$ and $R_1 = 50\%$, we adjusted the values of K_2 and R_2 to analyze their impact on the performance of HiMAP pruning. The experimental results are presented in Table 1. For short-text response generation tasks, such as ScienceQA and A-OKVQA, setting $K_2 = 8$ and $R_2 = 75\%$ effectively minimizes computational overhead while maintaining model performance. However, for long-text response generation tasks, such as Nocaps and Flickr30k, a more conservative configuration, $K_2 = 15$ and $R_2 = 75\%$, is necessary to ensure the model's performance remains uncompromised.

13.2. Effect of Pruning Module

Tab. 9 presents the results of ablation studies conducted on the pruning modules. It is evident that applying either the shallow-layer or deeper-layer pruning module individually can reduce computational overhead without compromising model performance. This demonstrates that both modules effectively accelerate model inference.

Model	SHL-PM	DPL-PM	TFLOPs	FLOPs Ratio	ScienceQA	A-OKVQA
LLaVA-v1.5-7B	✗	✗	2.98	100%	67.9	76.6
	✓	✗	1.56	54%	**68.3**	**77.1**
	✗	✓	1.78	34%	**68.1**	**77.2**
	✓	✓	0.73	24%	68.3	77.2
LLaVA-v.15-13B	✗	✗	5.81	100%	71.6	82.0
	✓	✗	3.09	53%	**71.8**	81.2
	✗	✓	1.73	30%	**72.0**	81.3
	✓	✓	1.36	23%	72.1	81.4

Table 9. **Ablation Study on Shallow-layer Pruning Module and Deeper-layer Pruning Module.**

14. Prompts for different tasks

Sci-VQA Dataset. In the Sci-VQA dataset, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

Sci-VQA Dataset

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**
Context: Select the best answer.
Which property do these three objects have in common?
A. shiny B. slippery C. opaque
Answer with the option's letter from the given choices directly.

ASSISTANT:

AOKVQA Dataset. In the AOKVQA dataset, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

A-OKVQA Dataset

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**
Analyse the image and choose the best answer for the following question:
What is in the motorcyclist's mouth?
Options: (A) toothpick (B) food (C) popsicle stick (D) cigarette
Output the letter of the correct answer.

ASSISTANT:

POPE Benchmark. In the POPE benchmark, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

POPE Benchmark

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**
Is there a cow in the image? Please just answer yes or no.

ASSISTANT:

Nocaps & Flickr30k Datasets. In the Nocaps and Flickr30k dataset, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

Nocaps & Flickr30k Datasets

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**
Provide a one-sentence caption for the provided image.

ASSISTANT: